4B

УДК 004.93'11

Е.В. Волченко
Институт информатики и искусственного интеллекта
ГВУЗ «Донецкий национальный технический университет», г. Донецк, Украина
Украина, 83050, г. Донецк, просп. Б. Хмельницкого, 84, LM@mail.promtele.com

Метод совместного построения решающих правил и выбора словаря признаков по взвешенным обучающим выборкам

E.V. Volchenko
Institute of Informatics and Artificial Intelligence
Donetsk National Technical University, Donetsk, Ukraine
Ukraine, 83050, Donetsk, 84 B. Khmelnitsky avenue

Method of Joint Construction of Decision Rules and Feature Selection Based on Weighted Training Samples

О.В. Волченко
Інститут інформатики і штучного інтелекту
ДВНЗ «Донецький національний технічний університет», м. Донецьк, Україна
Україна, 83050, м. Донецьк, просп. Б. Хмельницького, 84

Метод сумісної побудови розв'язувальних правил та словника ознак по зважених навчаючих вибірках

В работе предложен новый подход к решению задачи совместного построения решающих правил классификации и рабочего словаря признаков в обучающихся системах распознавания, основанный на использовании взвешенных обучающих выборок. Описаны метод w-GridDC формирования взвешенной обучающей выборки *w*-объектов, алгоритм *w*-MIEF построения рабочего словаря признаков на её основе и модифицированный метод *k*-ближайших соседей для выполнения классификации объектов. Приведены результаты экспериментальных исследований, подтвердившие эффективность предложенного подхода.
Ключевые слова: адаптивная система распознавания, обучающая выборка, *w*-объект, словарь признаков.

A new approach to solving the problem of joint constructing of classification decision rules and feature selection in the recognition training systems based on weighted training samples use is given in the article. W-GridDC method of forming a weighted training sample of w-objects, the algorithm w-MIEF of construction the feature set based on it, and a modified method of k-nearest neighbor for objects classification are described. Experimental results have confirmed the effectiveness of the proposed approach.
Key words: adaptive recognition system, training samples, w-object, feature selection.

У роботі запропонований новий підхід до вирішення задачі побудови розв'зувальних правил класифікації та робочого словника ознак у системах розпізнавання, що навчаються. Основою цього підходу є використання зважених навчальних вибірок. Наведено метод *w*-GridDC формування зваженої навчальної вибірки *w*-об'єктів, алгоритм *w*-MIEF побудови робочого словника ознак на її основі та модифікований метод *k*-найближчих сусідів для виконання класифікації об'єктів. Наведено результати експериментальних досліджень, що підтверджують ефективність запропонованого підходу.
Ключові слова: адаптивна система розпізнавання, навчальна вибірка, *w*-об'єкт, словник ознак.

Введение

Основными задачами, решаемыми при построении обучающихся систем распознавания, являются задачи построения эффективных решающих правил классификации и формирования рабочего словаря признаков [1]. На сегодняшний день одной из основных проблем при решении этих задач является большой объем исходных данных, которые должна обрабатывать система за выделенное время. Так, в задачах автоматической классификации текстов в новостных лентах и электронных библиотеках словарь признаков может состоять из тысяч ключевых слов, а обучающая выборка содержать десятки и сотни тысяч текстов [2].

Необходимость построения рабочего словаря признаков, состоящего в выборе оптимального набора наиболее информативных признаков из множества всех признаков априорного словаря, обусловлена значительными временными и емкостными затратами на измерение всех признаков классифицируемого объекта и выполнение классификации [1], [3], [4]. Задача построения решающего правила классификации состоит в формировании по всем объектам обучающей выборки выражения или алгоритма классификации распознаваемых объектов, обеспечивающего минимальную ошибку классификации [1], [3]. В теории построения систем распознавания эти задачи в большинстве случаев рассматриваются независимо друг от друга, хотя выбор алгоритма построения решающих правил в значительной мере зависит от набора признаков, которыми описываются распознаваемые объекты, а построенный словарь признаков оценивается качеством классификации по построенному решающему правилу.

При комплексном подходе к построению решающих правил классификации и формированию рабочего словаря признаков данную задачу, согласно [5], называют задачей комбинированного типа DX (построения решающего правила D в наиболее информативном подпространстве признаков X). Её сложность состоит в необходимости одновременного решения двух ключевых задач распознавания: построения эффективного решающего правила классификации, для которой в большинстве случаев увеличение количества признаков приводит к повышению эффективности классификации и минимизации словаря признаков для сокращения временных и емкостных затрат на выполнение классификации [6].

Наиболее известными алгоритмами решения задачи типа DX являются:

– алгоритмы CORAL и DW [1], основанные на переборе всех возможных словарей признаков и формировании на их основе решающих правил классификации;

– алгоритм FRiS-DX [6], основанный на выборе очередного варианта признакового подпространства и построения в нем решающего правила на основе алгоритма FRiS-Stolp, выделяющего подмножества эталонных объектов, на основании которых выполняется классификация.

Данные алгоритмы на основании анализа полной исходной информации выделяют некоторое признаковое подпространство, а затем выполняют анализ всех объектов для построения решающих правил. Учитывая большой объем обрабатываемых исходных данных, можно предположить, что такой подход потребует больших временных и вычислительных затрат из-за необходимости при выборе каждого нового признакового подпространства обрабатывать все исходное множество данных.

В данной работе на основе разработанных ранее алгоритмов [7], [8] предлагается новый подход к решению задачи типа DX, который состоит в предварительном сокращении исходной выборки путем перехода к сокращенным взвешенным выборкам *w*-объектов и дальнейшем построении оптимального словаря признаков и решающего правила классификации на их основе.

Целью данной работы является разработка метода совместного построения решающих правил классификации и рабочего словаря признаков на основе взвешенных выборок *w*-объектов в обучающихся системах распознавания.

Постановка задачи

В качестве исходных данных дано некоторое множество объектов $X=\{X_1, X_2, .., X_k\}$, представленное в виде объединения непересекающихся классов $X = \bigcup_{i=1}^{l} V_l$ и называемое обучающей выборкой. Каждый объект X_i из X описывается системой признаков, т.е. $X_i = \{x_{i1}, x_{i2}, \ldots, x_{in}\}$, и представляется точкой в линейном пространстве признаков, т.е. $X_i \in R^n$. Для каждого объекта X_i известна его классификация $y_i \in [1, l]$.

Решение задачи классификации некоторого объекта $X_s = \{x_{s1}, x_{s2}, \ldots, x_{sn}\}$ предполагает выполнение двух этапов обработки исходной выборки X для получения сокращенной по количеству объектов и их признаков взвешенной обучающей выборки и выполнение непосредственной классификации.

На первом этапе необходимо сформировать классифицированную взвешенную обучающую выборку *w*-объектов $X^W = \{X_1^W, X_2^W, \ldots, X_m^W\}$, $y_i^W \in V$, где $X_i^W = \{x_{i1}, x_{i2}, ..., x_{in}, p_i\}$, p_i – вес *i*-го *w*-объекта.

На втором этапе необходимо по выборке *w*-объектов построить сокращенный рабочий словарь из k $(k \le n)$ признаков при условии неухудшения эффективности классификации.

Построение выборки *w*-объектов по исходной обучающей выборке на основе алгоритма w-GridDC

Общим принципом построения сокращенных обучающих выборок *w*-объектов [8], [9] является выделение областей компактного расположения объектов одного класса в пространстве признаков и замена этого множества объектов одним *w*-объектом, вес которого характеризует количественные или топологические особенности найденного множества. Приведем далее обобщенное описание метода *w*-GridDC построения выборки *w*-объектов.

Идеей метода *w*-GridDC [8] является наложение сетки на признаковое пространство для формирования множества клеток, определение объектов выборки, принадлежащих каждой из клеток и их замена на *w*-объекты. Формирование объектов новой выборки выполняется только в случае принадлежности всех объектов клетки одному классу. Вес *w*-объектов определяется по количеству объектов исходной выборки, принадлежащих клетке.

Далее приведем пошаговое описание метода. Без потери общности получаемых решений применим стандартный для теории распознавания подход, заключающийся в рассмотрении двухклассовых систем.

Шаг 1. Формирование сетки. Рассчитывается шаг клетки *s* по формуле:

$$s = \left\lfloor 1 + \frac{\left(\sum_{i=1}^{n} (\max\{x_i\} - \min\{x_i\}) \right)^n * (\lfloor \ln(k) \rfloor - 1)}{n * \prod_{i=1}^{n} (\max\{x_i\} - \min\{x_i\})} \right\rfloor,$$

где $\lfloor ... \rfloor$ – оператор округления до ближайшего целого значения;

$\max\{x_i\}$ – максимальное значение i-го признака среди всех объектов выборки, $\min\{x_i\}$ – минимальное значение.

Выполняется разбиение признакового пространства R^n по каждому из n признаков на интервалы длиной s (наложение прямоугольной сетки), результатом которого является множество клеток G. Далее для каждого объекта выборки X определяется клетка, которой этот объект принадлежит.

Показано [8], что объект X_i принадлежит некоторой клетке G_j тогда и только тогда, когда каждое из значений его признаков входит в интервал значений соответствующих признаков данной клетки.

В результате формирования сетки и обработки объектов исходной обучающей выборки будут сформированы непересекающиеся подмножества X_{G_j} объектов, принадлежащих соответствующим клеткам G_j, $j = \overline{1,|G|}$.

Шаг 2. Формирование значений признаков *w*-объектов.

Возможны следующие варианты обработки содержимого клеток.

1. Если все объекты клетки принадлежат к одному классу, то значения признаков объекта новой выборки рассчитываются как координаты центра масс объектов этой клетки:

$$x_{jt} = \frac{1}{\left|X_{G_j}\right|} \sum_{X_i \in X_{G_j}} x_{it}, \ t = \overline{1,n}.$$

2. Если клетка не содержит ни одного объекта, то объект новой выборки не формируется.

3. Если клетка содержит объекты нескольких классов, то она делится на две равные по размеру клетки (поочередно вертикально или горизонтально) до тех пор, пока любая из клеток внутри начальной клетки не будет содержать объекты только одного класса. Далее по каждой из полученных клеток формируются объекты новой выборки (согласно случаям 1 и 2).

Классификация *w*-объекта определяется по классификации объектов, по которым он сформирован.

Шаг 3. Определение веса *w*-объектов. Вес *w*-объекта равен количеству объектов исходной выборки, принадлежащих клетке, т.е.

$$p_j = \left|X_{G_j}\right|.$$

В результате выполнения алгоритма будет получена новая взвешенная обучающая выборка *w*-объектов X^W.

Построение рабочего словаря признаков на основе взвешенных обучающих выборок *w*-объектов

Введение в описание объектов взвешенной обучающей выборки новой характеристики – веса – не позволяет эффективно использовать известные алгоритмы построения рабочего словаря признаков. Для решения задачи построения рабочего словаря признаков по взвешенным обучающим выборкам *w*-объектов в [7] предложен метод *w*-MIEF, описание которого приведем далее.

Основой метода *w*-MIEF является оценка индивидуальной информативности признаков по обобщенному критерию, включающему в себя отношение дискриминанта Фишера и коэффициент индивидуальной эффективности распознавания. Основной особенностью предлагаемого метода является использование в качестве исходных данных взвешенной выборки *w*-объектов.

Алгоритм *w*-MIEF состоит из следующих этапов.

1. Инициализация алгоритма (значений дискриминанта Фишера $fisher_i = 0, i = \overline{1,n}$ и среднего значения $fisherAvg = 0$).

2. Вычисление отношения дискриминанта Фишера для всех признаков априорного словаря

$$fisher_i = \frac{\sum_{j=1}^{l} P_j \cdot D_{ji}}{\sum_{j=1}^{l} \sum_{\substack{a=1 \\ a \neq j}}^{l} \left(P_j \cdot P_a \cdot (m_{ji} - m_{ai})^2 \right)},$$

где D_{ji} – дисперсия значений i-го признака по j-му классу,

m_{ji} – среднее значение i-го признака по j-му классу.

3. Вычисление степени покрытия классов признаками априорного словаря

$$\mathrm{cov}erage_{ji} = \frac{\sum_{a=1}^{S_i} featValItems_{ja}}{m_{ji} \cdot P_j},$$

где S_i – количество уникальных найденных значений признака x_i,

$featValItems_{ja}$ – количество объектов класса j с заданным значением i-го признака.

4. Вычисление взвешенной эффективности признаков

$$weightedEff_i = \frac{\sum_{j=1}^{l} featValWeight_j}{\sum_{a=1}^{k} p_a},$$

где $featValWeight_j$ – суммарный вес объектов класса j с заданным значением i-го признака,

p_a – вес *w*-объекта.

5. Выбор наилучших признаков:

1) с максимальной эффективностью по степени перекрытия классов:

$$BestCoverage = \max_{i=\overline{1,n}} \sum_{j=1}^{l} \mathrm{cov}erage_{ji},$$

2) с максимальной эффективностью по информативности:

$$BestFi = \max_{i=\overline{1,n}} weightedEff_i.$$

6. Формирование рабочего словаря признаков по параметрам максимальной эффективности по степени перекрытия классов и информативности признаков по следующим правилам:

1) эффективность текущего признака совпадает с эффективностью лучшего признака;

2) эффективность признака больше 0 и данный признак улучшает распознавание хотя бы одного класса;

3) эффективность признака больше 0 и распознавание хотя бы одного из классов лежит в пределах порога t, заданного пользователем.

Отметим, что наличие пользовательского порога t связно с тем, что возможны ситуации, когда признаки не улучшают распознавание хотя бы одного из классов, но, тем не менее, имеют эффективность, близкую к лучшей эффективности распознавания классов. Таким образом, если для рассматриваемого признака разница между лучшей и текущей эффективностью распознавания любого из классов находится в пределах $(0;t]$, то данный признак также считается информативным и включается в рабочий словарь признаков.

В результате работы алгоритма *w*-MIEF будет сформирован рабочий словарь, содержащий k ($k \le n$) признаков.

Выполнение классификации на основе взвешенной выборки *w*-объектов

Для классификации объектов на основе взвешенных обучающих выборок *w*-объектов будем использовать модифицированный алгоритм k-ближайших соседей [9], широко применяющийся при решении задач классификации в условиях неполных априорных данных. Выбор данного метода для классификации на основе взвешенной обучающей выборки основывается на результатах исследований [10], согласно которым он будет показывать высокую эффективность классификации при использовании сокращенной обучающей выборки.

Модификация алгоритма *k*-ближайших соседей в данном случае будет состоять в использовании метрики, позволяющей определять близость между объектами взвешенной обучающей выборки и классифицируемым объектом X_s [10]:

$$d_W(X_i^W, X_s) = \frac{\sqrt{\sum_{o=1}^{n}(x_{io} - x_{so})^2}}{p_i}.$$

Для определения классификации X_s найдем *k*-ближайших к нему *w*-объектов каждого из классов и отнесем к тому классу, суммарное расстояние до *w*-объектов которого минимально.

Результаты экспериментальных исследований

Для оценки эффективности предложенного подхода был проведен ряд экспериментальных исследований. В качестве исходных данных были использованы выборки объектов двух классов размером 1000 – 5000 объектов при 20% пересечении областей классов в пространстве признаков, содержащих 10-100 признаков распознавания. Для генерации значений признаков использовался нормальный и равномерный законы распределения. Также для экспериментальных исследований были использованы наборы данных репозитория UCI [11].

В качестве критерия оценки эффективности классификации использовалась частота неверных классификаций. Количество «ближайших соседей» было выбрано равным 10% размера обучающей выборки *w*-объектов.

При анализе результатов предложенного подхода к решению рассматриваемой задачи были получены следующие результаты:

1) размер взвешенных выборок *w*-объектов составил в среднем 2,3% размера исходных обучающих выборок;

2) количество признаков, включенных в рабочий словарь, составило 40% - 55% исходного количества признаков;

3) частота неверной классификации объектов тестовой выборки модифицированным методом *k*-ближайших соседей по выборке *w*-объектов с использованием рабочего словаря признаков уменьшилась в среднем на 3,7% по сравнению с частотой неверной классификации методом *k*-ближайших соседей по исходной выборке.

4) эффективность предложенного подхода увеличивалась с увеличением размера исходных обучающих выборок.

Таким образом, результаты тестовых испытаний подтвердили эффективность предложенного подхода к совместному построению решающих правил классификации и рабочего словаря признаков на основе взвешенных выборок *w*-объектов.

Выводы

В данной работе предложен новый подход к решению задачи типа DX построения эффективного решающего правила классификации с определением минимального словаря признаков, основанный на переходе к взвешенным обучающим выборкам *w*-объектов. Описаны метод *w*-GridDC формирования взвешенной обучающей выборки *w*-объектов, алгоритм *w*-MIEF построения рабочего словаря признаков на её основе и модифицированный метод *k*-ближайших соседей для выполнения классификации объектов. Анализ предложенного подхода показал сходимость составляющих его методов, их низкую временную сложность, корректность обработки объектов исходной выборки. Отличительной особенностью данного подхода является использование сокращенной взвешенной обучающей выборки, что позволяет существенно сократить временные и емкостные затраты на построение словаря признаков и выполнение классификации объектов при сохранении начального уровня эффективности работы системы.

Литература

1. Загоруйко Н.Г. Прикладные методы анализа знаний и данных / Загоруйко Н.Г. – Новосибирск : Издательство института математики, 1999. – 270 с.
2. Pal K.S. Pattern recognition: from classical to modern approaches / K.S. Pal, A. Pal – Calcutta : World scientific, 2001. – 612 p.
3. Theodoridis S. Pattern Recognition / S.Theodoridis, K. Koutroumbas. – San Diego : Academic Press, 2008. – 823 p.
4. Загоруйко Н.Г. Проблема выбора в задачах анализа данных и управления / Н.Г. Загоруйко, Г.С. Лбов // Сиб. журн. индустр. матем. – № 3:1. – 2000. – С. 101-109.
5. Загоруйко Н.Г. Методы распознавания и их применение. / Загоруйко Н.Г. – М. : Сов. радио, 1972. – 206 с.
6. Борисова И.А. Использование FRiS-функции для построения решающего правила и выбора признаков (задача комбинированного типа DX) / И.А. Борисова, В.В. Дюбанов, Н.Г. Загоруйко, О.А. Кутненко // Труды Всероссийской Конференции «Знания-Онтологии-Теории» (ЗОНТ-07), Новосибирск, 2007. – Том I. – С. 37-44.
7. Волченко Е.В. Метод w-MIEF построения рабочего словаря признаков на основе взвешенных обучающих выборок / Е.В. Волченко, В.С. Степанов // Вісник Національного технічного університету «Харківський політехнічний інститут». Збірник наукових праць. Тематичний випуск: Нові рішення в сучасних технологіях. – Харків : НТУ «ХПІ», 2012. – № 18. – С. 26-33.
8. Волченко Е.В. Сеточный подход к построению взвешенных обучающих выборок w-объектов в адаптивных системах распознавания / Е.В. Волченко // Вісник Національного технічного університету «Харківський політехнічний інститут». Збірник наукових праць. Тематичний випуск: Інформатика і моделювання. – Харків : НТУ «ХПІ», 2011. – № 36. – С. 12-22.

9. Волченко Е.В. Метод построения взвешенных обучающих выборок в открытых системах распознавания / Е.В. Волченко // Доклады 14-й Всероссийской конференции «Математические методы распознавания образов (ММРО-14)», Суздаль, 2009. – М. : Макс-Пресс, 2009. – С. 100-104.
10. Волченко Е.В. О способе определения близости объектов взвешенных обучающих выборок / Е.В. Волченко // Вісник Національного технічного університету «Харківський політехнічний інститут». Збірник наукових праць. Тематичний випуск: Інформатика і моделювання. – Харків : НТУ «ХПІ», 2012. – № 15. – С. 12-20.
11. Merz C.J. UCI Repository of machine learning datasets / C.J. Merz, P.M. Murphy // Information and Computer Science University of California, Irvine, CA, 1998. – Режим доступа: http://www.ics.uci.edu/~mlearn/databases

Literatura

1. Zagorujko N.G. Prikladnye metody analiza znanij i dannyh. Novosibirsk: Izdatelstvo instituta matematiki, 1999. 270 s.
2. Pal K. S. Pattern recognition: from classical to modern approaches. Calcutta: World scientific. 2001. 612 p.
3. Theodoridis S. Pattern Recognition. San Diego: Academic Press. 2008. 823 p.
4. Zagorujko N.G. Problema vybora v zadachah analiza dannyh i upravleniya. Sib. zhurn. industr. matem. №31. 2000. S. 101-109.
5. Zagorujko N.G. Metody raspoznavaniya i ih primenenie. M.: Sov. Radio. 1972. 206 s.
6. Borisova I.A. Trudy vserossijskoj konferencii "Znaniya-Ontologii-Teorii" (ZONT-07). Novosibirsk 2007. T. I. S. 37-44.
7. Volchenko E.V. Visnyk nacionalnogo tehnichnogo universytetu "Harkivskyj politehnichnyj instytut". Zbirnyk naukovih prac'. Tematychnij vypusk: Novi rishennya v suchasnih tehnologiyah. Harkiv: NTU "HPI". 2012. № 18. S. 26-33.
8. Volchenko E.V. Visnyk nacionalnogo tehnichnogo universytetu "Harkivskyj politehnichnyj instytut". Zbirnyk naukovih prac'. Tematychnij vypusk: Informatyka i modelyuvannya. Harkiv: NTU "HPI". 2011. № 36. S. 12-22.
9. Volchenko E.V. Doklady 14 Vserossijskoj konferencii "Matematicheskie metody raspoznavaniya obrazov MMRO-14". Suzdal. 2009. M.: Maks-Press. 2009. S. 100-104.
10. Volchenko E.V. Visnyk nacionalnogo tehnichnogo universytetu "Harkivskyj politehnichnyj instytut". Zbirnyk naukovih prac'. Tematychnij vypusk: Informatyka i modelyuvannya. Harkiv: NTU "HPI". 2012. № 15. S. 12-20.
11. Merz C.J. Information and Computer Science University of California. Irvine. CA. 1998. http://www.ics.uci.edu/~mlearn/databases

RESUME

E.V. Volchenko

Method of Joint Construction of Decision Rules and Feature Selection Based on Weighted Training Samples

The problem of joint constructing of decision rules of classification and feature selection in training recognition systems is considered in the work. The main problem is the need to analysis of large volume of input data. Existing methods for its solution are considered, features of their implementation are defined.

The weighted training samples using to reduce the amount of computation is proposed, where each object is formed by the set of objects in the original sample. The grid method w-GridDC of forming a weighted training set of objects was described. Algorithm w-MIEF of construction of feature set using a weighted sample, based on an assessment of information content of features of weight w-objects is given. The modified method of k-nearest neighbors for objects classification using a weighted sample is proposed.

The effectiveness of the proposed approach is confirmed by experiments on test and real data. It is shown that this approach can significantly reduce the time and space costs for the construction of the feature set and decision rules while maintaining an entry-level of the system efficiency.

Статья поступила в редакцию 04.07.2012.